Desayunos ASTIC

Almacenamiento y análisis de grandes volúmenes de información

POR MAOLE CEREZO
REDACTORA JEFE DE BOLETIC
FOTOS AITOR DIAGO

Evento patrocinado por

EMC² where information lives

SYBASE

Para ofrecer su visión sobre el almacenamiento y análisis de grandes volúmenes de información, las empresas EMC y Sybase participaron en un desayuno organizado por ASTIC que fue conducido por Carlos Maza. La importancia del tema, tal y como recordó el Vicepresidente de ASTIC "no se escapa a nadie, ya que la Ley de Acceso ha hecho que se escalen los volúmenes de información, pasando de almacenar alguna y estructurada, a toda la generada en las organizaciones".

En estos tiempos, "hay un fortísimo desarrollo tecnológico en lo que se refiere a temas de almacenamiento, nuevas tecnologías para disco, inteligencia del software de base de datos..., que va muy relacionado con la calidad y la disponibilidad de los servicios. La Ley nos obliga a dar servicios 24x7 y hay muchos temas de actualidad ligados al almacenamiento, como pueden ser los centros de respaldo, sistemas de back up, bases de datos distribuidas...".

La realidad que se vive en las AAPP, tal y como explicó Carlos Maza, es que "disponemos de un gran conjunto de bases de datos, pero en su explotación se le ha dado una prioridad al trabajo operativo, en línea, al corto plazo y, con frecuencia, no se han agotado las posibilidades del análisis de explotación de esa información, con el objeto de mejorar la políticas publicas de la administración".

Antes de entrar a fondo en el tema de debate, las compañías hicieron unas breves presentaciones. Así, Javier Sánchez, Comercial Senior de EMC, recordaba como "nacimos con el objetivo de ayudar a nuestros clientes a custodiar de la forma más óptima la información. Cuando empezamos a trabajar en EMC, lo que más nos preocupaba era consolidar todos los sistemas distribuidos en distintos servicios o servidores, en sistemas centrales, para dotarlos de la mayor seguridad posible y optimizar el uso del almacenamiento que se hace". Hoy en día, prosigue "este planteamiento quizás no tenga tanto sentido, por el fuerte crecimiento de los volúmenes de información que manejamos. El crecimiento de información viene dado, sobre todo, por los datos no estructurados, y esto ha hecho que hablemos, en vez de consolidación, de estratificación y de automatizar esa estratificación. Se trata de alinear los datos con el nivel de servicios que requerimos de ellos, situándolos en aquella infraestructura que suponga el menor coste posible: dar el nivel de servicio que se requiere, al menor coste".

Para ilustrar su argumento sacó a colación el sistema de correo, "que crece de forma exponencial, y el uso más crítico se concentra en volumen de datos muy pequeño. Lo mismo sucede con los ficheros de usuarios, los más críticos, con respecto del volumen de información que analizamos, son muy pequeños. Un 80% de ella es no crítica, no exige gran nivel de servicio, sólo un 20 es la que necesitamos

tener viva, requiere rendimientos muy buenos".

Explicó como en la Administración están trabajando en diversas soluciones de archivado de información, utilizando "sistemas de almacenamiento de gama más alta, con discos de nueva tecnología para la información viva que exige un alto nivel de servicio, y sistemas de archivado que no ofrecen gran rendimiento, para datos históricos, a los que se accede con menos frecuencia"

Desde hace más de un año en EMC se está trabajando con "una tecnología nueva para el estrato del on- line más rabioso, tecnología de discos flash que elimina de los sistemas de almacenamiento los tiempos mecánicos. Su coste es más elevado, por lo que se utiliza para la información más viva". La propuesta más novedosa de la compañía es la nueva gama de productos de almacenamiento Symmetrix V-Max "una revolución basada en una tecnología greed, donde se comparten la memoria, la CPU, y los canales de entrada salida dentro de la máquina". Los sistemas de almacenamiento Symmetrix V-Max se han diseñado de acuerdo a la nueva tecnología Virtual Matrix Architecture, que interconecta los múltiples módulos V-Max. Cada uno de estos módulos contiene su propia CPU redundante, memoria y conectividad de servidor/disco. Esta arquitectura permite una alta escalabilidad y fiabilidad, además de niveles extremos de tolerancia contra los fallos.

Para el archivo de información "a largo plazo", EMC propone "plataformas de archivado con dispositivos basados en direccionamiento por contenido, que permiten que un administrador sea capaz de gestionar un tamaño enorme de Teras". Nos dan la oportunidad "no solo disminuir los costes de almacenamiento (puro coste por Gb.), sino también los de gestión de estas plataformas". Entre los beneficios del archivado automático, el más inmediato, "es la reducción de costes. El coste por Tb. de estos sistemas de archivado es menor que el de los discos flash o el de los sistemas tradicionales de alto rendimiento". Esta arquitectura "permite reducir los costes de infraestructuras de servidores y los medioambientales del CPD, a la vez que mejora el servicio". Tal y como explica Javier Sánchez, "en el archivado, podremos abordar mayores crecimientos de información de una forma más óptima, sin necesidad de ir incorporando cada vez más infraestructura". Otras ventajas

El desayuno congregó a 25 personas entre socios y representantes de EMC y Sybase

Vicente Moncho. Director de Marketing de Sybase

Carlos Maza

Carmen Cabanillas

operativas son "la mejora de los mecanismos de replicación y operativas en el caso del back up y en la restauración de la información". La apuesta de EMC es "alinear el ciclo de vida de nuestra información con las infraestructuras que hay por debajo, lo que nos permitirá el ahorro de costes, lo que estamos haciendo en la Administración".

Sybase es una compañía que nació en 1984 creando una base de datos. Sybase®, Inc. (NYSE: SY) es, tal y como señaló su Director General en España, Joaquín Berenguer, "la mayor compañía de software del mundo especializada en gestionar y movilizar información corporativa. Sybase cuenta con un historial de más de 20 años como líder en tecnología, y los datos más importantes de los sectores de comercio, finanzas, gobierno, atención médica y defensa de todo el mundo se ejecutan en Sybase". En los tiempos de incertidumbre que corren, la compañía se muestra en plena forma, "los ingresos por licencias, el beneficio operativo y los ingresos totales de Sybase han crecido más del 10% en el segundo trimestre de 2009". Entre algunos de sus hitos históricos caber citar que "Sybase es la primera empresa del mercado en ofrecer bases de datos relacionales cliente/servidor, proporcionando al proyecto sobre el genoma humano licencias para la primera generación de bases de datos relacionales cliente/servidor (1988) y la primera empresa en ofrecer tecnología de replicación abierta (1990).

En banca y telecomunicaciones, la mayoría de las firmas "son nuestros clientes". Su división de movilidad da servicio a gigantes internacionales como "Coca Cola o Pepsi Cola" y cuenta con referencias en la AAPP en todo el mundo, desde departamentos de seguridad, institutos de estadística, aduanas, censo, sanidad, ministerios de defensa..., donde se utiliza grandes volúmenes de información. Entre algunas referencias importantes en su cartera de clientes de España figuran la Agencia tributaria o el Ejército del Aire, que utiliza los sistemas de replicación de Sybase para coordinar entre sí todas sus bases operativas. Tal y como destacó Joaquín Berenguer, "la obtención en 2002 del premio a la mejor base de datos móvil ("Mobility Award for Best Database")" que obtuvo su producto Sybase iAnywhere avalan su trabajo.

De primera mano, los presentes en el encuentro, tuvieron la oportunidad de conocer cómo Sybase maneja los grandes volúmenes de información. Si de lo que se trata es de que "la consulta sea libre, y de que los tiempos de respuesta sean de segundos, no de minutos, partimos de un servidor analítico, un software que está especializado en una arquitectura basada en columnas —diferente de la que serviría para almacenar información referente a DNIs, y sus datos, para la que sirven las bases de datos relacionales basadas en filas— al mismo

tiempo que se produce una compresión de datos, para que la arquitectura tenga menos necesidad de recursos del sistema". Su pócima magistral "son las columnas. Con el almacenamiento dirigido a columnas vamos verticalmente al dato que nos interesa". Con la tecnología de Sybase, "se produce un ahorro de tiempo, gracias a nuestra manera de almacenamiento de la información. El resultado final: consumo menos energía y almacenamiento".

Experiencias

Cómo podemos saber ¿qué tenemos en un gran volumen de información no estructurada, procedente de orígenes diversos, con el fin de dividirla en cajitas? ¿Tenéis alguna experiencia o algún proyecto al que referiros?, planteó Esther Fernández, de la Oficina Nacional de Investigación del Fraude de la Agencia Tributaria.

La alternativa que ofrece EMC se remite a su producto Documentum, con una suite de soluciones que, entre otras cosas, permite capturar información ya digitalizada, como la que está en formato papel, y extraer los datos más importantes para hacer una clasificación automática. Tratadas "con un gestor documental, cualquiera que tengamos en nuestra organización, podría resolverse lo que propones. Con Captiva, se puede abordar este tipo de problemática, ya que reconoce formatos tipo imagen o cualquier otro fichero ofimático. EMC *Captiva Family* transforma documentos, faxes y fuentes electrónicas de datos que son críticos en contenido listo y apropiado para que se procesen las aplicaciones del negocio".

Javier Bustillo, Director Comercial para la AAPP de EMC abundó argumentando que, como fabricante, "damos respuesta a este problemática desde el punto de vista de la infraestructura. Utilizando discos en estado sólido, cuando el número de operaciones a realizar es muy elevado, necesitas infraestructura que te permita tener millones de operaciones de entrada salida, y para ello, hay sistemas de almacenamiento con discos de estado sólido que te permiten un rendimiento más alto, y un tiempo de respuesta más pequeño".

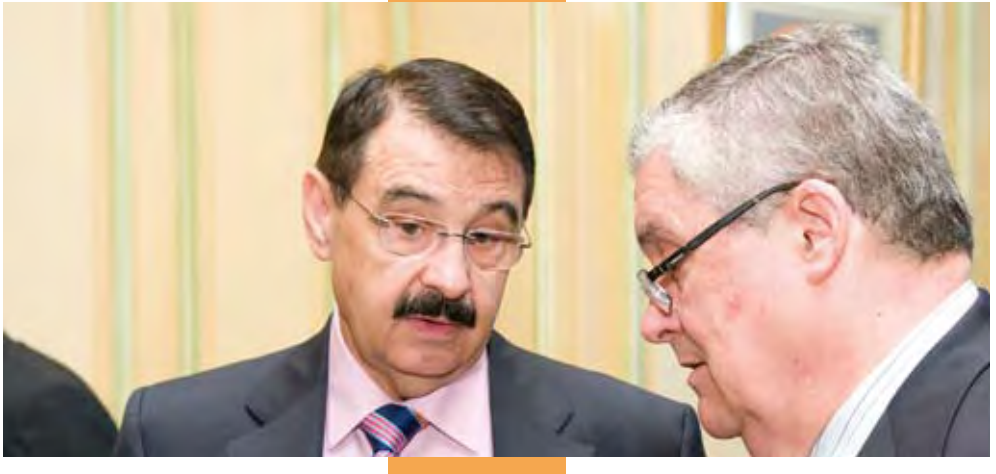
Por su parte, Joaquín Berenquer, se refería al motor de base de datos IQ, en su versión más reciente 12.7., diseñado específicamente para entregar resultados más rápidos en soluciones de inteligencia empresarial analítica de misión crítica, almacenes de datos y generación de reportes, combinando velocidad y agilidad, con un bajo coste total de propiedad. Una vez, "contamos con esa información en nuestra base de datos (sean imágenes, vídeo, ficheros office o mails…), a partir de ahí, para identificar el formato hay diversas herramientas en el mercado. Con la tecnología de que dispone la agencia tributaria, podría archivar los 25 teras a los que te refieres en el IQ

Carmen García Roger

Esther Sánchez

Felipe Jusdado

Fernando Martín



Fernando Ruiz y Guillermo Searle

Carmen Cabanillas

que ya tiene". IQ "es una base de datos a la que le pedimos lo mismo que a una base de datos operacional orientada a filas, su diferencia está en la forma de analizar los datos"

Datos Muertos

Los grandes volúmenes de información no estructurada que actualmente ya disponen las AAPP irán creciendo inevitablemente, y las bolsas de datos muertos se convierten en un caballo de batalla con el que tienen que lidiar las organizaciones, además de con los datos en línea. Por motivos históricos, y de acuerdo a la Ley de Archivos, habrá que almacenarlo todo, con lo que ello supone en cuanto al coste de licencias, espacio en las bases de datos etc. Carlos Maza, Presidente de ASTIC incidió en la importancia de este hecho, pidiendo soluciones a las empresas presentes: ¿qué tecnología tenéis para detectar esas bolsas de datos muertos?

Tener información, no estructurada, en base de datos es "un problema", apuntó Javier Sánchez. Lo que proponen en EMC es que "cuando la información que se almacena es no estructurada, se guarden en base de datos los índices que permiten tener ordenada esta información. Pero la información, per se, tiene que estratificarse y almacenarse en contenedores de información no estructurada. Tenemos soluciones y plataformas a nivel de aplicativo con las que estratificar la información con tipología diversa". Es una realidad que "el 80% de los datos que manejamos es información no estructurada, en tal caso, no podemos almacenarla en base de datos".

Si bien, en opinión de los técnicos de las empresas "la tendencia de almacenar en base de datos es mejor y seguirá haciéndose de esta manera", el Director General de Sybase coincidiendo con EMC en la necesidad de "la estratificación de la información", propone "otra manera de hacerlo". Como "los datos no sabemos si están o no muertos y no tenemos la seguridad de que necesitemos emplearlos a diario, se ha de contar con la misma base de datos que contenga toda información en línea, a la vez que se ha de disponer de un sistema inteligente capaz de buscar la información, esté donde esté". Pero como lo que no se quiere es que "moleste día a día", la solución está en "una misma tecnología estratificada en diferentes entornos".

Cuando la información muerta se encuentra en entornos como el forense, los datos que se manejan son de vital importancia. Recordó Guillermo Searle, funcionario que en la actualidad trabaja en la Agencia Tributaria, con años de experiencia en el Ministerio de Justicia, que "normalmente, hay una gran dilación entre el momento en el que se produce la información —que en la actualidad se está grabando en CD y puede ser visual, auditiva y en diversos lenguajes— y cuando tienen lugar los juicios. Ello requiere herramientas como las que facilita la ingeniería informática que se dedica al Data Mining. Esta información se termina almacenando en un soporte magnético, un soporte direccionable... Tiene que existir un procedimiento de expurgo para decidir que información es la que se puede eliminar o no, pero ha de ser una labor absolutamente consciente".

Ignacio Cudeiro, Director de Boletic, trabajando actualmente en el Ministerio de Ciencia e Innovación, puntualizó como se puede observar la información "tanto desde el punto de vista de su importancia, y desde la de su frecuencia de acceso, contemplando cómo está evolucionando en cuanto a su actualidad". Para éste, el problema se encuentra en que "parte de nuestra información no sabemos de qué tipo es y en que sistema almacenarla, dependiendo de la necesidad de frecuencia de acceso". La pregunta "que todos hacemos es: ¿Qué soluciones técnicas hay para analizar el grado de importancia de la información que tenemos, el tipo y la frecuencia de acceso para, en función de ello, decidir bajo que plataforma física almacenarla con el fin de no tener problemas ante peticiones de restauración inmediata?. Ello, teniendo en cuenta, que no vamos a poder estructurarla y meterla en discos de alto rendimiento".

"Dependiendo de la fuente de datos donde tengamos la información, será más o menos fácil", respondió Javier Sánchez. El escenario ideal es "cuando toda la información la contenemos en un gestor documental. Documentum es capaz de, dependiendo de las políticas del servicio que se configuren, mover la información de un sitio a otro, de un disco de fibra a un dispositivo tipo Centera. Pero cuando nos encontramos en un sistema de fichero, con un volumen muy grande de información, es más complicado". Porque "carecemos de la información sobre el contenido de esos datos, y aquí lo que podemos hacer es archivar en base a la información que nos viene del propio sistema operativo, por ejemplo, en función del tamaño, del tipo de fichero, de cuando

Javier Bustillo. Director de AAPP de EMC

Javier Sánchez. Comercial Senior de EMC

Joaquín Berenguez. Director General de Sybase

José Antonio García

José Luis Gil, Ignacio Cudeiro y Pablo Burgos

Juan Fernando Muñoz

ha sido archivado por última vez....” Cuando “hablaba de archivado automatizado, me refería a que todas estas operaciones se hacen de forma desatendida, nosotros lo que solo hacemos es escribir las normas por las cuales se mueve la información, y este movimiento es hacia el archivado y hacia la inversa”.

Qué las necesidades de almacenamiento de datos crecen exponencialmente, es algo con lo que todos los presentes estuvieron de acuerdo, y ante esta realidad, los directivos de la Administración no cesan en buscar soluciones. Si bien son varias las ofertas en el mercado, cambiar no resulta fácil, tal y como planteó Blas Cordero, del Ministerio de Asuntos Exteriores, sacando a colación un reciente estudio sobre los sistemas de gestión de bases de datos relacionales en el que se decía que “a los usuarios les costaba mucho cambiar de una base de datos a otra”. Por ello, quiso saber “las razones para cambiar de nuestro sistema de base de datos actual al de Sybase”.

Nuevamente Joaquín Berenguer se refirió a dos perspectivas desde las que abordar la cuestión “las filas y las columnas”. En lo que se refiere a las filas, “Sybase tiene más éxito cuando hay muchas transacciones y son cortas. Aquí, en los sectores de banca y telecomunicaciones, ganamos. La mayoría de las empresas de Wall Street funcionan con Sybase. En los análisis de grandes volúmenes de información, nos remitimos a las columnas, en donde no hay nadie más que nosotros”.

Reducción o incremento

¿Pueden estas tecnologías animar al incremento de información no controlada?, preguntó José Manuel Pacho, de Patrimonio del Estado. Si bien durante el desarrollo del desayuno una idea recurrente fue la de trabajar de una manera más sostenible, no pareció convencer a todos que los ahorros de energía fueran fruto directamente del empleo de las tecnologías tratadas. Para éste, “una primera prevención que habría que observar sería identificar los datos, y posteriormente, discriminar cuáles tendrían que seguir siendo almacenados

José Antonio García

de los que no". Con las tecnologías "estamos almacenando datos que antes no se almacenaban, y ello puede llevar a la no sostenibilidad"

En cuanto a la transición de tecnologías o la adquisición de nuevas propuestas del mercado, Pacho puso sobre la mesa el hecho de que "cada vez estas tecnologías de almacenamiento, que tiempo atrás estaban categorizados, son más intangibles, con una componente de servicio creciente. Ello complica mucho la contratación en la Administración, partiendo de la base de que, incluso, resulta muy difícil determinar la naturaleza del mismo contrato". En su opinión, hoy "la normativa va por detrás del mercado".

La tesis sobre el "falso" ahorro de energía fue apoyada por Vicente Moncho, Director de Marketing de Sybase, reconociendo que "a veces, la tecnología puede ser un catalizador para almacenar más información de la necesaria, y conseguir el efecto contrario al que buscamos en cuanto al ahorro de energía se refiere". Pero esta necesidad, está obedeciendo según el directivo, "a un cambio de paradigma, en el que los sistemas de información se está convirtiendo en algo imprescindible en la tarea diaria. Tenemos que responder a unas nuevas necesidades que nos obligan a incorporar un nuevo ciclo de vida para la información". Y al hilo de todo esto, al hablar de información viva o muerte

—prosigue— "distinguiría entre la que es para el puro transaccional o para el análisis de esa información a posteriori. De cualquier forma, viva o muerta, necesitamos tener esa información y, por ello, se ha de guardar".

A un cambio tecnológico importante, en lo que se refiere al almacenamiento, que va a ofrecer al mercado más capacidad a un precio más reducido, se refirió Fernando Martín, del Tribunal de Cuentas, mostrando su interés por conocer más detalles. A la vez, recalcó como él "guarda todo, la información viva o muerta y por duplicado. Javier Sánchez, le recordó como "cuando aparecieron los discos flash, de 73 Gb. a unos precios elevados, supusieron una cierta revolución". Hoy "ya existen de 400 Gb., y los precios han bajado". Los informes de los gurús de EMC vaticinan que, "en futuro próximo, la tecnología mecánica ciber chanel será sustituida por los discos flash, aunque permanezca

Víctor Pérez.
Director de AAPP de Sybase

José Luis Gil, Miguel Ángel Orellana y Miguel Ángel Rodríguez

José Manuel Pacho

Leonor Torres

Rocío Montalbán

Blas Cordero

una tecnología de bajo coste". Por su parte, Javier Bustillo, comentó como "en entornos universitarios norteamericanos se está investigando para que el siguiente portador físico de almacenamiento se logre a nivel más pequeño, con mayor nivel de integración y pueden construirse memorias más rápidamente".

Por las "técnicas de duplicación" se interesó Carmen Cabanillas, del Ministerio de Industria a quien Javier Sánchez, de EMC le precisó como "la duplicación, en sí, es un concepto no un producto y dependiendo de cómo se aplique se obtendrán unas ventajas u otras". Hoy en día, "la aplicamos en diferentes productos, unos son los sistemas de archivados que hacen duplicación a nivel de fichero, con el fin de que éstos no se guarden n veces". La duplicación se está aplicando mucho también "a los ámbitos de back up, pues allí se tiende a guardar la misma información con frecuencia, en librerías virtuales, donde nos permite, en muchos casos, eliminar las cintas dado que los discos cada vez son más grandes. También en entornos virtuales, lugar en el que tenemos uno de los principales problemas con el back up..."

La experiencia de Juan Fernando Muñoz en el Ministerio de Sanidad, donde se trabaja con diferentes organizaciones: comunidades autónomas, hospitales..., sirvió para cerrar el encuentro. El directivo compartió como trabajan "a nivel nacional, europeo, con diferentes productos y distintas formas de organizar la información, con lo cual lo único que puedes hacer es introducir unos procedimientos, a través de los cuales, recopilas la información, la indexas a nivel macro entre organizaciones. Acuerdas los procedimientos para buscar elementos que permitan la interoperatividad".

Carlos Maza concluyó señalando que, a lo largo del encuentro, se había estado hablando de "una tecnología en alza, un sector donde se está invirtiendo, y en el que los costes se mantienen". A la vez, invitó a los fabricantes a tener en cuenta "el aspecto de la fiabilidad, seguridad y accesibilidad a las bases de datos". "Tenéis que estar especialmente sensible a ello porque el usuario agradece mejoras de tiempos de respuesta, necesitamos tener garantizada disponibilidad del dato y fiabilidad del servicio".